《术语标准化与信息技术》，2008 年第四期，总第 52 期，4-9.

一个新兴的术语学科 -- 计算术语学

冯志伟

摘要：术语在科学技术文献中分布很广，术语的自动处理和识别对于科学技术文献的分析、识别和生成至关重要。本文介绍了术语学中的一个新兴学科—计算术语学，分别介绍了术语的发现、术语的充实、术语的受控标引、术语的自由标引等问题。

关健词：计算术语学，术语发现，术语充实， 受控标引，自由标引。

A new scientific domain in terminology

- Computational Terminology

Feng Zhiwei

Abstract: Terms are pervasive in scientific and technical documents; their automatic processing and identification are the crucial issues for any application dealing with the analysis, understanding, generation or translation of such documents. This paper introduces a new scientific domain in terminology - computational terminology, it presents term acquisition, term enrichment, term controlled indexing and term free indexing.

Key words: computational terminology, term acquisition, term enrichment, controlled indexing, free indexing.

近年来，在术语学的研究中，开始引进自然语言的计算机处理的方法和技术，出现了“计算术语学”[1]（computational terminology）这样的学科。1998 年的计算语言学国际会议 COLING-ACL’98 上，组织了世界上第一次计算术语学的讨论会（First Workshop on Computational Terminology），这次讨论会首次使用的“计算术语学”这个学科名称。这次讨论会讨论的问题主要有：

- 如何抽取术语以满足信息检索的需要；
- 如何抽取术语以便使用双语语料库来进行翻译；
- 如何进一步完善和原有术语抽取的工作（例如，如何建立概念层级网络，如何搜索语义信息或概念信息）。

[1] D. Bourigault, Ch. Jacquemin, Marie-Claude L’Homme, Recent Advances in Computational Terminology, John Benjamins Publishing Company, 2001.

1998 年的这次讨论会成为了计算术语学发展的催化剂，从此，计算术语学便成为一个新兴的术语学的学科，活跃在当代科学技术的百花园中，并且一天天地成熟起来，初步具备了系统的理论和有效的方法，值得我们特别地关注。在“计算术语学”这个名称出现 10 年之前，我国冯志伟在 1988 年就注意到术语的自动处理问题，他在德国夫琅禾费研究院（Fraunhofer Institute）使用计算机对汉语的词组型术语进行了自动结构分析，是国际上最早进行计算术语学研究的学者之一[2]。计算术语学的研究主要包括术语结构的自动剖析、术语的自动发现、术语的自动标引等。 本文主要介绍术语的术语的自动发现和术语的自动标引。

在自然语言的计算机处理的诸多领域中，都离不开术语，例如，机器翻译（machine translation）目前主要是翻译专业性的文献，术语的自动处理与机器翻译系统的译文质量有密切的关系；此外，信息检索（information retrieval）、信息抽取（information extraction）、文本分类（text classification）的运算的基本单位都是单词型术语或词组型术语，也离不开术语的自动处理。

术语是自然语言处理中的一种特殊的词汇数据，与语言中一般的普通词汇不同，术语大多数都是由多个单词组成的词组型术语，它们对于科学技术的发展特别敏感，时时刻刻随着科学技术的发展而发展。在术语的发展过程中，它们不断地丰富，不断地充实，不断地变化，术语的语义也在不断地转移，旧的术语消失了，新的术语产生了。在这样的情况下，术语数据库需要经常地维护，不断地用新的术语充实原来的内容，有时甚至需要重建，以反映科学技术的日新月异发展的要求。这样，术语的发现（term detection）或术语的获取（term acquisition）就成为了术语自动处理的一个重要内容。术语发现可以进一步分成两个类型：如果在术语发现中不依赖初始的术语数据，那么，这样的术语发现叫做“初始术语发现”（initial term acquisition）；如果在术语发现中要使用初始的术语数据，那么，这样的术语发现叫做“原有术语充实”（term enrichment）。

在文本自动处理中，术语的使用与术语的自动辨识（term recognition）是紧密联系在一起的。术语的自动辨识主要研究如何进行术语的自动标引（automatic indexing）。在自然语言处理中，为了便于信息的存取，文本文献总是要使用单词表或词组表，因此，有必要在文本文献中进行术语的自动标引（automatic indexing of terms），然后根据自动标引的结果，使用计算机来自动地生成单词型术语表或词组型术语表。由于术语是科学技术知识在自然语言中的结晶，术语能够浓缩地表示特定的科学技术领域中的主要概念，它们可以被看成是文本内容的抽象描述，文本文献经过术语的自动标引之后，就能大体上反映出其内容。因此，在文本自动处理中，术语的自动标引是非常重要的。

根据在标引时是否依赖初始的术语数据，术语的自动标引也可以分为两个类型：如果在术语标引中不依赖初始的术语数据，那么，这样的术语标引叫做“自由标引”（free indexing）；如果在术语标引中要使用初始的术语数据作为参照，那么 ，这样的术语标引叫做“受控标引”（controlled indexing）。

总起来说，术语自动处理可以这样来分类：如表 1 所示：

---

[2] Feng Zhiwei, Analysis of Chinese Terms in Data Processing, Report in Fraunhofer Institute, 1988, Stuttgart.

表 1 术语自动处理的四个主要领域

	不依赖于初始术语数据	依赖于初始术语数据
术语发现	初始术语发现	原有术语充实
术语辨识	自由标引	受控标引

下面我们介绍国外的术语发现研究和术语标引研究情况[3]。

首先介绍“术语发现”的研究。发现候选术语的方法基本上分为符号法（symbolic approach）和统计法（statistical approach）两种。符号法根据术语（主要是名词词组）的句法描述来发现候选术语；统计法根据词组型术语中组成成分的互信息（Mutual Information）来发现术语，组成成分之间的互信息越大，它们组成术语的可能性也就越大。

(1) 基于语法的术语发现方法：例如，在 1994 年， Lauriston 在 TERMINO 系统中提出了一种基于语法的术语发现方法，这种方法要对文本进行剖析，利用文本中的单词和句法线索（lexical and syntactic clues）来发现术语[4]。剖析模型的操作顺序如下：

a. 预处理：首先对文本进行过滤，除去对于术语发现无用的那些形式特征(虚词，停用词)；
b. 剖析并抽取术语：
   - 形态分析；
   - 名词短语剖析；
   - 术语生成。
c. 交互式术语数据库的构建和管理：给用户提供友好的界面，把前面步骤中抽取出来的术语构建成术语数据库。

(2) 句法模式与机器学习到的选择限制相结合的方法：例如，在 1996 年，D. Bourigault 研制的术语自动处理工具 LEXTER[5]。LEXTER 使用带标记的语料库，语料库中的标记有词汇特征的标记和句法模式的标记两种，这个工具有一个可视化的界面，可用来确认并组织从带标记的语料库中抽取出来的术语。

a. 最大名词短语的分离：LEXTER 可使用分离规则，从最大名词短语（maximal noun phrase）中把可能性最大的术语边界分离出来。例如，在法语的最大名词短语中，过去分词与介词结合而成的组合很可能是术语的边界，在法语最大名词短语 les clapets situés sur les tubes d’alimentation （位于进气管上的阀门）中，situés sur 是术语的边界，把整个名词短语分离为 les clapets（阀门） 和 les tubes d’alimentation（进气管） 两部分，这两部分分别是两个不同的术语。 其中，“situés sur”是句法模式，这个模式的使用取决于动词的选择限制，而动词的选择限制是通过内置的机器学习程序从语料库中自动地学习得到的。

[3] Christian Jacquemin, Spotting and Discovering Terms through Natural Language Processing, The MIT Press, 2001.
[4] A. Lauriston, Automatic recognition of complex terms: problems and the TERMINO solution, *Terminology*, 1(1), 147-170, 1995.
[5] D. Bourigault, LEXTER: a natural language tool for terminology extraction, *Proceedings of the 7th EURALEX International Congress*, 771-779, 1996.

b. 把最大名词短语分解成候选术语：确定边界之后，最大名词短语被分离为两个部分，通过后处理，最后由人来判定这些候选术语，并把确认后的术语加入到术语数据库中。例如，从最大名词短语 les clapets situés sur les tubes d'alimentation 中，把术语 les clapets 和 术语 les tubes d'alimentation 自动地抽取出来，作为候选术语，加入到术语数据库中。又如，在法语中，pylône à haute tension （高压电线架）的结构是：N + prep + N + Adj，经过最大名词短语分离之后，把 haute tension（高压电）作为候选术语提取出来，加入到术语数据库中。
c. 最后，还可以根据这些候选术语在句法位置上的相似程度，把它们组织起来。例如，法语中的 vanne motorisés（电动门）、vanne pneumatique （气动门）、vanne d'alimentation（进气门）都有共同的中心词 vanne，就把它们组织起来，形成一组有关系的候选术语。
d. 这些进入术语数据库的候选术语，由专家做最后的审定，确定为正式的术语，充实了原有的术语。

(3) 句法模式与统计过滤相结合的方法：例如，在 1996 年，Daille 研制的 ACABIT 是一个把句法模式与统计过滤结合起来的术语研究工具[6]。ACABIT 获取候选术语的步骤如下：

a. 语言规则过滤（linguistic filtering）：根据术语结构的语言学规则，使用有限状态转移网络发现候选术语，在英语中，主要考虑三种模式的术语：Adj + N, N + N, N + Prep + N。由这三种模式扩展而形成的变体，也可以作为候选术语的筛选范围。例如，satellite transit network （N + N + N）可以看成是由 N + N 模式扩展而成的，multiple satellite links（Adj + N + N）可以看成是由模式 Adj + N 和模式 N + N 扩展而成的。
b. 统计排序（statistical ranking）：使用某些统计方法，对前面的步骤筛选出来的候选术语进行排序。例如，计算候选术语的“对数似然度”（log-likehood ratio），根据计算结果对于候选候选术语排序，得出在统计意义上可能性最大的术语。

(4) 抽取搭配信息的方法：例如，在 1993 年，Smadja 研制的 Xtract 是一个专门用于抽取搭配关系的工具[7]。Xtract 的重点不是关心术语本身，而是关心术语在意义上的可搭配性。只有那些在语义上可以搭配的词语才可以算做候选术语（例如，stock trader[存货商人]，last selloff [最后的存货]在语义上是可以搭配的）。
(5) 非语言学的方法：例如，Enguehard 和 Pantera 在 1995 年研制的术语提取工具 ANA[8]。ANA 是独立于具体语言的术语自动抽取工具，它包括两个模块：

a. 预熟悉模块（familiarization module）：使用预熟悉模块来确定三类词语：

- 停用词语表（stop list）：停用词通常是一些频度很高的词语，这些词语都不具有专业性。
- 种子术语表（set of seed terms）：使用人工从语料库中选出反映专业概念的术语作为种子术语（seed term），构成种子术语表。

[6] B. Daille, Study and implementation of combined technique for automatic extraction of terminology, In *The balancing Act: Combining Symbolic and statistical Approaches to language*, MIT Press, 49-66, 1996.
[7] F. Smaja, Retrieving collocation from text: Xtract, *Computational Linguistics*, 19(1), 143-177, 1993.
[8] C. Enguehard and L. Pantera, Automatic natural acquisition of a terminology, *Journal of Quantitative Linguistics*, 2(1), 27-32, 1993.

- 结构词语表（set of scheme words）：这些结构词语一般是介词或限定词之类的虚词，它们在语料库中往往与种子术语一起出现。

b. 发现模块（discovery module）：使用机器自动学习中的“自举”（bootstrap）方法，一步一步地扩充从预熟悉模块中得到的种子术语的规模，从而发现更多的术语。

在用于术语发现的这五种方法中，前两种方法（TERMINO，LEXTER）不使用统计，假定文本中符合条件的全部词语都是候选术语，哪怕只出现一次的“罕用词语”（hapax legomenon），只要它们符合条件，也都在候选术语的考虑范围之内。这两种方法是非统计的方法。使用这样的非统计方法时，术语的判定要由用户来进行，需要给用户提供交互工具，以便用户对于候选术语进行选择。后面三种方法都要使用统计来进行过滤或排序，在这样的情况下，考虑候选术语出现的上下文环境就显得非常重要了，因为统计的数据需要在具体的文本或语料库中才可以计算出来，离开了具体的文本或语料库，不可能进行任何的统计，当然也就不可能发现术语了。

术语辨识主要是做术语的自动标引。传统的自动标引主要使用“词口袋”（bag-of-words）的方法，这种方法只是简单地把所标引的单词直接地与它们所在的文本联系起来，基本上不考虑这些单词的语言结构信息。这是“词口袋”技术的缺点。实际上，在术语的自动标引时，应当保持术语中单词的顺序，还要反映出术语的结构以及术语中单词之间的依存关系，这时，“词口袋”技术就显得不足了。为了反映单词的语言结构信息，需要对于术语进行自动剖析。术语自动剖析的深度取决于具体的需要，可以进行浅层的句法剖析，也可以进行比较深层的句法分析。根据自动剖析的深度，术语的自动标引可以分为基于浅层句法剖析的自动标引和基于深层句法剖析的自动标引。基于浅层句法剖析的自动标引使用的标引技术有文本简化（text simplification）、基于窗口的关键词识别（window-based keyword recognition）等。基于深层句法剖析的自动标引使用的标引技术有基于依存关系剖析的自动标引和基于转换剖析的自动标引。下面介绍三种简单的术语自动标引方法。

(1) 文本简化方法：例如，在 1983 年，Dillon 和 Gray 研制的 FASIT 系统就使用了文本简化的方法[9]。FASIT 的自动标引分两步：

a. 标注与模式匹配：FASIT 首先使用后缀规则和不规则后缀的特例表对于文本进行形态分析，对有关的词语进行词类标注，然后把分析得到的带有词类标记的文本与表示术语结构的句法模式（例如，N, N + N, Proper-noun + N 等）相匹配，得到有关术语的句法模式的标引。

b. 标引合并：使用文本简化技术，把得到的句法模式标引进行合并，合并步骤如下：

-- 删除停用词（如，介词，连接词，普通名词）；

-- 词根还原；

-- 词序重组。

这样 ，便可以得到带有句法模式的术语标引。

[9] M. Dillon and A. S. Gray, FASIT: a fully automatic syntactically based indexing system, *Journal of American Society for Information Science*, 34(2), 99-108, 1983.

(2) 名词词组的歧义消解方法：例如，在 1991 年，Evans 研制的 CLARIT 系统[10]，把自然语言处理中的形态分析技术、浅层剖析技术和统计过滤技术结合起来，对于名词短语进行歧义消解。首先，对文本进行形态分析，使名词短语术语中的单词得到没有歧义的词类标记。然后对所得到的带有词类标记的名词短语术语进行句法剖析，得到候选的名词短语结构（不考虑结构歧义）。例如，名词短语 the redesigned R3000 chips from DEC（来自 DEC 公司的重新设计 R3000 的芯片）经过这样的剖析之后，得到

$$[\text{the}]_{\text{Det}}\ [\text{redesigned R3000}]_{\text{PreMod}}\ [\text{chips}]_{\text{Head}}\ [\text{from DEC}]_{\text{PostMo}}$$

其中，Det 表示限定词，Head 表示中心词，PreMod 表示前修饰语，PostMod 表示后修饰语。

剖析得到的候选术语再根据统计特征进行排序。

在使用 CLARIT 时是不考虑结构歧义的，因此，标引的结果还需要进一步使用基于语料库的技术进行结构消歧，得到没有结构歧义的标引。

（3）用于自动标引的句法剖析方法：有一些研究者使用句法剖析器从文本中抽取名词短语术语。剖析时术语的语法关系的表示方法主要有两种：一种是基于结构成分的分析方法，一种是基于依存关系的分析方法。

a. 基于结构成分的分析方法：例如，在 1995 年，Strzalkowski 研制的 TTP 剖析器可以产生出词组型术语的树形结构，在树形结构中，表示出中心词（head）和它有关的论元（argument）[11]。例如，名词短语 the former Soviet president （前苏联的总统）被分析为如下的树形结构：

$$[_{\text{NP}}\ [_{\text{N}}\ \text{president}]\ [_{\text{T-pos}}\ \text{the}]\ [_{\text{Adj}}\ [\text{former}]]\ [_{\text{Adj}}\ [\text{Soviet}]]]$$

TTP 剖析器是根据比较全面的英语语法来设计的，使用了“语言串语法”（Linguistic String Grammar）的理论，语法范畴主要来自《牛津高级英语学习词典》（Oxford Advanced Learner Dictionary）。

由 TTP 剖析器分析得出的词组型术语，可以用来从文本中自动地生成术语标引。由于经过标引后的这些术语都带有句法结构的信息，对于机器翻译、信息检索等自然语言处理是非常有用的。

在 1990 年，Metzler 设计了成分对象剖析器 COP（Constituent Object Parser），这个剖析器只使用二元的依存关系信息，由于树形结构中的支配关系具有传递性，一个具有 n 层依存关系的树形结构可以转换成具有 n-1 层的二叉树形结构，这样，所有的树形结构都可以变成二元的树形结构。例如，small liberal arts college for scared junior（为胆小的少年办的小型的自由艺术学校）可以被分析为如下的树形结构：

[*[small *[liberal *[arts *college]]][for *[scared *junior]]]

---

[10] D. A. Evans and C. Thai, Noun-phrase analysis in unrestricted text for information retrieval, *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics* (ACL'96), 17-24, 1996.

[11] T. Strzalkowski, Natural language Information Retrieval, *Information Processing and Management*, 31(3), 397-417, 1995.

其几何形状为：

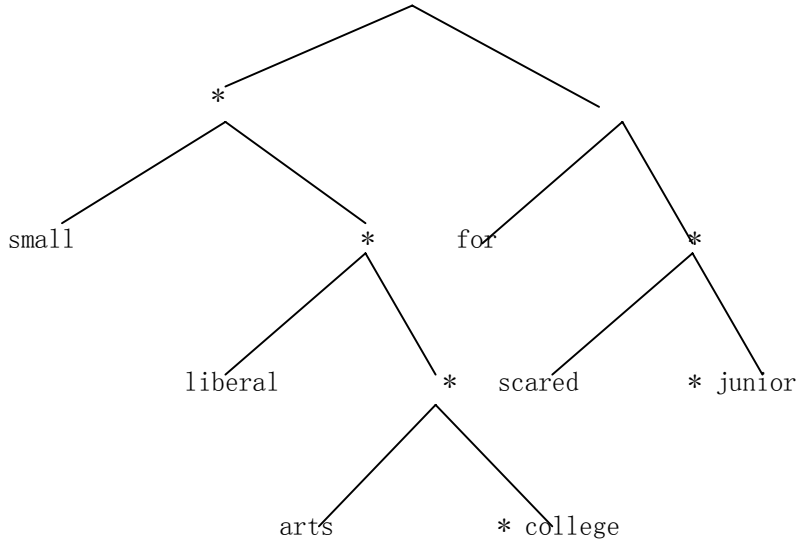

图 1　表示二元关系的二叉树

其中的每一个子树都是二元的，标有*号的子树是中心语，没有*号的成分是附加语，根结点上没有加任何的标记，子树[for *[scared *junior]]是修饰 college 的，也不代任何的标记。从这个二叉树中可以看出，中心语标记*是具有继承关系的，它们可以由下层传递到上层。

b. 基于依存关系的分析方法：在 1988 年，Schwarz 研制了 COPSY 系统，这个系统使用法国语言学家特尼耶尔（L. Tesnière）提出的“依存语法”（dependency grammar）[12],对名词短语术语进行自动剖析，剖析的结果要表示出名词短语术语中的依存关系。例如，problems of fresh water storage and transport in containers or tanks （用集装箱或水箱储存和运输的新鲜水的问题）经过 COPSY 剖析之后，可以得到如下的依存关系：

fresh → water

water → storage → problem　　water → transport → problem

container → storage　　container → transport

tank → storage　　tank → transport

其中，“→”表示“依存于”，例如，fresh → water 表示 fresh 依存于 water。这些依存关系是根据名词短语术语中单词之间的结构特性建立起来的，是依存分析的结果。

(4) 术语变体的识别方法：一个术语往往会存在若干个不同的变体（variation），因此，在术语的计算机自动处理中，还要研究术语变体的识别问题。1999 年，Jacquemin 研制了 FASTR

[12] 关于依存语法，可参看冯志伟《现代语言学流派》（修订本），陕西人民出版社，1999 年。

系统[13]，使用结构转换与词汇关系结合的方法来识别术语变体。术语的词汇关系可以反映在形态的联系上（例如，具有相同的词根的术语在形态有联系），也可以反映在语义的联系上（例如，同义术语，反义术语）。FASTR 可以识别出 malignancy in orbital tumours （眼窝肿瘤的恶性）是 malignant tumour（恶性的肿瘤）的变体，因为 malignancy（恶性） 和 malignant（恶性的）在形态上相关，它们都包含词干 malignan-，而且，malignancy in orbital tumours 的结构模式为 N + Prep + Adj + N，这个模式与 FASTR 系统定义过的名词短语模式 N + Prep + Adj + N 相匹配，据此可以判断它是一个词组型术语。

术语的变体有三类：

- 句法变体：这种变体只与句法有关；
- 形态变体：这种变体与形态结构的转换和音位的变化有关；
- 语义变体：这种变体只与语义有关。

FASTR 是为受控标引而研制的。这个系统首先输入一个权威性术语表，把它转换成可计算的数据，并自动生成这些术语的候选变体。然后再把这些候选变体与语料库中的数据相比较，最后检索出真正的术语变体。

上面介绍的都是单语言的术语自动处理，下面我们介绍双语言的术语自动处理。

双语言的术语自动发现一般要分两步走。第一步是术语抽取，在双语言的语料库中分别进行术语自动抽取，找出每一种语言中的术语；第二步是术语对齐（alignment），找出在不同语言之间术语的对应关系。

双语言的语料库中术语的对齐有不同的方法。例如，Gaussier 的方法是，先进行句子的对齐，然后再在已经对齐的句子中进行术语对齐[14]，这是一种先处理大的语言单位，再处理小的语言单位的“从大到小”方法。但是，Hull 则提出了不同的方法。他先进行单词型术语对齐，再进行术语抽取，最后进行词组型术语的对齐。单词型术语的对齐和词组型术语的对齐都使用了无回溯的“贪心算法”（greedy algorithm）[15]。这是一种先处理小的语言单位，后处理大的语言单位的“从小到大”方法。

计算术语学是一个新兴的术语学的学科，这个学科的出现，反映了信息网络时代对于术语学研究的新要求，是信息网络时代对于术语学的挑战，值得我们密切关注。关于术语的自动发现和术语的自动辨识方法，今后我们还可以研究如下问题；

- 建立大规模的专业语料库，开展专业语料库的研究，进行基于语料库的语义标注研究和语义关系自动获取的研究。
- 研究专业语料库构建的新技术。
- 在大规模的专业语料库中，获取更多的语义学资源和形态学资源，以便为术语或术语变体的自动发现提供可靠的数据。

---

[13] C. Jacquemin, Syntagmatic and paradigmatic representation of term variation, *Proceedings of the 37th Annual Meeting of the Association of Computational Linguistics (ACL'99)*, 341-348, 1999.

[14] E. Gaussier, Flow network models for word alignment and terminology extraction from bilingual corpora, *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98)*,444-450, 1998.

[15] D. Hull, Automating the construction of bilingual terminology lexicons, *Terminology*, 4(2), 225-244, 1997.

- 把基于规则的方法、基于统计的方法以及机器学习的方法结合起来，研究术语发现和术语辨识的新的“混合方法”（hybrid solution）[16]。
- 对专业语料库进行加工，使它带有更加丰富的信息，使普通的“上下文”（context）变成“富语境”（rich context），使语料库中的上下文更具有解释性和说明性，把一般上下文中的文本信息和富语境中包含的结构信息结合起来，进行术语的发现和辨识。
- 建立更加完善的交互界面，以便专业人员更方便地对候选术语进行人工判定。

计算术语学这个新兴的学科正在逐步成熟，我们应当关心这个学科的发展，并开展相应的研究，进一步丰富我国术语学的研究内容，推动我国术语学研究的现代化进程。

[16] Feng Zhiwei, Hybrid Approaches for Automatic Segmentation and Annotation of Chinese Text Corpus，*International Journal of Corpus Linguistics*, Vol. 6 (Special issue), 2001.